Improve the Evaluation of Fluency Using Entropy for Machine Translation Evaluation Metrics

Hui Yu[†] **Xiaofeng Wu**[‡] **Wenbin Jiang**[†] **Qun Liu**[‡†] **Shouxun Lin**[†]
[†]Key Laboratory of Intelligent Information Processing
Institute of Computing Technology, Chinese Academy of Sciences
[‡]ADAPT Centre, School of Computing, Dublin City University

Abstract

The widely-used automatic evaluation metrics cannot adequately reflect the fluency of the translations. The n-gram-based metrics, like BLEU, limit the maximum length of matched fragments to n and cannot catch the matched fragments longer than n, so they can only reflect the fluency indirectly. METEOR, which is not limited by n-gram, uses the number of matched chunks but it does not consider the length of each chunk. In this paper, we propose an entropy-based method, which can sufficiently reflect the fluency of translations through the distribution of matched words. This method can easily combine with the widely-used automatic evaluation metrics to improve the evaluation of fluency. Experiments show that the correlations of BLEU and METEOR are improved on sentence level after combining with the entropy-based method on WMT 2010 and WMT 2012.

1 Introduction

Automatic machine translation (MT) evaluation plays an important role in the evolution of MT. It not only evaluates the performance of MT systems, but also provides guidance for the improvement of MT systems (Och, 2003).

The automatic MT evaluation metrics can be classified into three categories: lexicon-based methods (Papineni et al., 2002; Snover et al., 2006; Lavie and Agarwal, 2007; Chen and Kuhn, 2011; Chen et al., 2012), syntax-based methods (Liu and Gildea, 2005; Owczarzak et al., 2007; Chan and Ng, 2008; Zhu et al., 2010; Mehay and Brew, 2007) and semantic-based methods (Lo et al., 2012), according to the employed information type. Most of the lexicon-based metrics obtain the similarity between the reference and hypothesis based on n-gram, such as BLEU (Papineni et al., 2002) and NIST(Doddington, 2002). BLEU obtains the score by a geometric mean of the n-gram precisions and a length-based penalty. NIST is closely related with BLUE but uses the arithmetic mean instead of geometric mean. For these metrics, the maximum length of matched fragments is limited to n, so they cannot catch the matched fragments longer than n. Some metrics which are not limited by n-grams relieve this problem, such as METEOR (Lavie and Agarwal, 2007). METEOR uses the Fmean of unigrams and a penalty. The penalty in METEOR is related to the number of matched chunks[1]. When the number of chunks in two sentence are the same, METEOR doesn't distinct them. The syntax-based metrics obtain the similarity by comparing the syntactic structures of two trees, and they cannot reflect the fluency directly. Semantic-based metrics, such as MEANT (Lo et al., 2012) which uses semantic role labeling (SRL) to match the predicate and arguments, mainly obtain the semantic information and do not consider the fluency.

In this paper, we propose an entropy-based method which can not only exploit the chunks with the maximum matched length but also reflect the difference between the lengths of the chunks. This method can easily combine with the widely-used automatic evaluation metrics to improve the evaluation of fluency. In the experiments, the new method is used to combine with BLEU and METEOR, and the sentence level correlations of

[1]The words in each chunk are in adjacent positions in the hypothesis, and are also mapped to unigrams that are in adjacent positions in the reference.

BLEU and METEOR are improved on WMT 2010 and WMT 2012.

2 Entropy-based Method

In this section, we introduce entropy and the entropy-based method (ENT) which can reflect the fluency of translations.

2.1 Entropy

Entropy is a measure of the uncertainty in a random variable. Shannon denoted the entropy H of a discrete random variable x with possible values $x_1, x_2, ..., x_n$. The entropy is defined as Formula (1) (Shannon, 2001).

$$H(X) = -\sum_{i=1}^{n} P(x_i) log_2 P(x_i) \qquad (1)$$

$P(x_i)$ is the probability of x_i showing up in the stream of characters. The more decentralized of the values $x_1, x_2, ..., x_n$, the higher of the entropy H(X). So the entropy can reflect the distribution of the values of variable x.

2.2 Entropy-based Method

In the automatic evaluation of machine translation, entropy can reflect the distribution of matched words. A lower entropy corresponds to a more concentrate distribution of matched words which represents a more fluent hypothesis. On the contrary, a higher entropy corresponds to a more disperse distribution of matched words, which represents a less fluent hypothesis. So the entropy-based method can reflect the fluency of translations sufficiently by the distribution of the words.

An example (a reference and three hypotheses) is shown as follows.

- ref: There are books on the desk
- hyp1: **There are books** in that **desk**
- hyp2: **There are** table **on the** book
- hyp3: **There are** table **on** book **the**

The matched words are in bold. hyp1, hyp2 and hyp3 can all match four words, but the distribution of the four words are different. The matched words are in two chunks for hyp1 and hyp2, and three chunks for hyp3. A smaller number of chunks represents a more concentrated distribution of the matched words, and corresponds to a more fluent hypothesis. From this point of view, hyp1 and hyp2 are better than hyp3. hyp1 has the same number of chunks as hyp2 but the number of the matched words in the two chunks is (3, 1) for hyp1 and (2, 2) for hyp2. hyp1 is considered to be more fluent than hyp2.

The details of the ENT are represented in following three steps. First, we obtain the matched words through the alignment of reference and hypothesis. The alignment is derived using Meteor Aligner[2]. The matched words are considered to be in a chunk if they are continuous and appear in the same order in both reference and hypothesis. Second, the entropy of chunks is calculated using Formula (2).

$$H = -\sum_{i=1}^{c} \frac{l_i}{L} log(\frac{l_i}{L}) \qquad (2)$$

l_i is the length of the ith chunk. c is the number of the chunks. L is the total number of the matched words. In the last step, the final score of ENT is achieved by Formula (3). To obtain a score within scope (0,1), an exponential function is used. We use $-H$ instead of H in the formula to ensure that a higher score of ENT represents a more fluent translation.

$$ENT = \alpha^{-H \times LP}, \quad \alpha \in (1, 1.5) \qquad (3)$$

LP, a length penalty, is calculated by Formula (4). l_h is the length of hypothesis. l_r is the length of reference.

$$LP = \beta^{|\frac{l_h}{l_r} - 1|}, \quad \beta \in (1, 2) \qquad (4)$$

Using Formula (3), the scores in the above example can be obtained as follows.

$$LP_{hyp1} = LP_{hyp2} = LP_{hyp3} = \beta^{|\frac{6}{6} - 1|} = 1$$

$$ENT_{hyp1} = \alpha^{-(-(\frac{3}{4} log \frac{3}{4} + \frac{1}{4} log \frac{1}{4})) \times 1} \approx \alpha^{-0.24}$$

$$ENT_{hyp2} = \alpha^{-(-(\frac{2}{4} log \frac{2}{4} + \frac{2}{4} log \frac{2}{4})) \times 1} \approx \alpha^{-0.30}$$

$$ENT_{hyp3} = \alpha^{-(-(\frac{2}{4} log \frac{2}{4} + 2 \times \frac{1}{4} log \frac{1}{4})) \times 1} \approx \alpha^{-0.45}$$

We can see that $ENT_{hyp1} > ENT_{hyp2} > ENT_{hyp3}$. Accordingly, the quality of hyp1 is obviously better than hyp2, and hyp2 is better than hyp3. So the entropy-based method can distinct these situations well.

[2] `http://www.cs.cmu.edu/~alavie/METEOR/`

The alignment of reference and hypothesis is derived only using *exact* match for ENT. We can also use linguistic resources to get the alignment, such as stem(Porter, 2001), synonym (Wordnet[3]) and paraphrase. In this case, we name the new method as ENTplus (ENTp).

3 Combine Entropy-based Method with Other Metrics

The new entropy-based method can effectively measure the fluency of a sentence. Most of the current metrics are good at the measure of accuracy, so we combine the entropy-based method with the widely-used automatic evaluation metrics to further improve the performance of these metrics. In this section, we take BLEU and METEOR as examples to introduce the combination method, but the entropy-based method can combine with most of the widely-used evaluation metrics.

3.1 Combine Entropy-based Method with BLEU

BLEU is a widely-used automatic evaluation metric owing to its simplicity and effectiveness. BLEU is calculated by Formula (5) (Papineni et al., 2002).

$$BLEU = exp(\sum_{n=1}^{N} w_n log p_n) \times BP \quad (5)$$

$$BP = \begin{cases} 1 & if\ c > r \\ e^{(1-r/c)} & if\ c \leq r \end{cases} \quad (6)$$

In Formula (5), the first part is a geometric mean of the n-grams precision where p_n is the precision of n-gram, and the second part is a length-based penalty as shown in Formula (6). There is also a length penalty in ENT. So we remove the part of length penalty in ENT when combining ENT with BLEU (Formula 7). The experience value of α is 1.05.

$$\text{BLEU}_{ENT} = BP \times exp(\sum_{n=1}^{N} w_n log p_n) \times \alpha^{-H} \quad (7)$$

[3] http://wordnet.princeton.edu/

3.2 Combine Entropy-based Method with METEOR

METEOR is calculated by Formula (8), in which Pen is calculated by Formula (9).

$$\text{METEOR} = Fmean \times (1 - Pen) \quad (8)$$

$$Pen = x1 \cdot (\frac{\#chunks}{\#unigrams\_matched})^{x2} \quad (9)$$

The first part in Formula (8) is the fmean of unigrams. The second part is related with the number of chunks. METEOR doesn't consider the length of each chunk, so it cannot reflect the situation that two hypotheses have the same number of matched unigrams and the same number of chunks, but different lengths for each chunk. We use ENT instead of $1 - Pen$ to reflect the above situation, and the final score can be computed in Formula (10). The experience value of α and β are 1.5 and 1.12 respectively.

$$\text{METEOR}_{ENT} = Fmean \times \alpha^{-H \times LP} \quad (10)$$

4 Experiments

To compute the correlation with human judges on sentence level, Kendall's rank correlation coefficient τ is employed. A higher value of τ means a better ranking similarity with the human judgments. τ is calculated as follows.

$$\tau = \frac{count_{con\ pairs} - count_{dis\ pairs}}{count_{total\ pairs}}$$

$count_{con\ pairs}$ is the count of concordant pairs. $count_{dis\ pairs}$ is the count of discordant pairs.

4.1 Data

In order to verify the effectiveness of ENT, we carry out the experiments on WMT 2010 and WMT 2012. There are four language pairs including German-to-English (de-en), Czech-to-English (cz-en), French-to-English (fr-en) and Spanish-to-English (es-en), which are all derived from WMT 2010 with 2034 sentences and WMT 2012 with 3003 sentences. The number of translation systems for each language pair is showed in Table 1.

Data	Metrics	cz-en	de-en	es-en	fr-en	ave
WMT10	BLEU	0.2554	0.2748	0.2805	0.2197	0.2576
	BLEU+ENT	0.2565	0.2730	0.2822	0.2211	0.2582
	BLEU+ENTp	**0.2643**	**0.2823**	**0.3010**	**0.2368**	**0.2711**(+1.35)
WMT12	BLEU	0.1567	0.1840	0.1938	0.1999	0.1836
	BLEU+ENT	0.1660	0.1907	0.1940	0.2060	0.1892
	BLEU+ENTp	**0.1732**	**0.1989**	**0.2052**	**0.2208**	**0.1995**(+1.59)

Table 2: Sentence level correlations of BLEU, BLEU+ENT and BLEU+ENTp on WMT 2010 and WMT 2012. The last column gives the average scores of the four language pairs.

Data	Metrics	cz-en	de-en	es-en	fr-en	ave
WMT10	METEOR	0.3292	0.3585	0.3283	0.2710	0.3218
	METEOR+ENTp	**0.3354**	**0.3593**	**0.3586**	**0.2923**	**0.3364**(+1.46)
WMT12	METEOR	0.2124	0.2748	0.2493	0.2506	0.2468
	METEOR+ENTp	**0.2153**	0.2730	**0.2585**	**0.2539**	**0.2502**(+0.34)

Table 3: Sentence level correlations of METEOR and METEOR+ENTp on WMT 2010 and WMT 2012. The last column gives the average scores of the four language pairs.

data	cz-en	de-en	es-en	fr-en
WMT2010	12	25	15	24
WMT2012	6	16	12	15

Table 1: The number of translation systems for each language pair on WMT 2010 and WMT 2012.

4.2 Experiment Results

The correlations of BLEU[4] are the results of 4-gram with smoothing option. According to the different methods of obtaining the chunks, we try two configurations, BLEU+ENT and BLEU+ENTp. BLEU+ENT is only using the exact match. BLEU+ENTp is using some resources which are stem, synonym and paraphrase. The correlations of METEOR are obtained from the released data of WMT 2010 (Version 1.27) and WMT 2012 (Version 1.48) with task option rank. We only do the experiment using outside resources (METEOR+ENTp), because METEOR also uses the outside resources. [5]

The sentence level correlations of the four language pairs and the average scores are shown in Table 2 and Table 3. In Table 2, BLEU+ENT is better than BLEU on both WMT10 and WMT12, but the result is only improved a little compared with BLEU. The reason is that the alignment is not good enough when only using exact match. When using stem, synonym and paraphrase, the result has a significant improvement of 1.35 on WMT 2010 and 1.59 on WMT 2012 respectively when comparing with BLEU. The number of reference is limited, and linguistic resources can enrich the reference, so ENTp can get better performance than ENT.

From Table 3, we can see that METEOR+ENTp has a significant improvement (1.46 on average) on WMT 2010, while the improvement on WMT 2012 (0.34 on average) is not as much as on WMT 2010. The METEOR version on WMT 2012 optimizes the parameters on the data of WMT 2009 and WMT 2010. We didn't tune the parameters after combing METEOR with entropy-base method, so the improvement is not very significant.

In all, when combining the entropy-based penalty with the widely-used automatic evaluation metrics, such as BLEU and METEOR, the performance can be improved, which proves the effectiveness of the entropy-based method.

5 Conclusion and Future Work

In this paper, we use entropy to reflect the fluency of the translation, and propose an entropy-based method. When combing the entropy-based method with the widely-used automatic evaluation metrics, such as BLEU and METEOR, the performances of these metrics are improved.

[4] ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v13a.pl

[5] Interested readers can find the source code of ENT and ENTp from https://github.com/YuHui0117/AMTE/tree/master/ENTFp.

One purpose of automatic evaluation metrics is to improve the quality of machine translation systems. So, in the future, we will use the combination of entropy-based method and widely-used metrics in the tuning process to improve the translation quality, such as MERT (Minimum Error Rate Training) (Och, 2003).

Acknowledgements

This work is supported by National Natural Science Foundation of P. R. China under Grant Nos. 61379086, 61602284, 61602285, 61602282 and Shandong Provincial Natural Science Foundation of China under Grant No. ZR2015FQ009. Qun Liu's work is partially supported by the Science Foundation Ireland (Grant 13/RC/2106) as part of the ADAPT Centre at Dublin City University.

References

[Chan and Ng2008] Yee Seng Chan and Hwee Tou Ng. 2008. Maxsim: A maximum similarity metric for machine translation evaluation. In *Proceedings of ACL-08: HLT*, pages 55–62.

[Chen and Kuhn2011] Boxing Chen and Roland Kuhn. 2011. Amber: A modified bleu, enhanced ranking metric. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 71–77, Edinburgh, Scotland, July. Association for Computational Linguistics.

[Chen et al.2012] Boxing Chen, Roland Kuhn, and George Foster. 2012. Improving amber, an mt evaluation metric. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, WMT '12, pages 59–63, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Doddington2002] George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145. Morgan Kaufmann Publishers Inc.

[Lavie and Agarwal2007] Alon Lavie and Abhaya Agarwal. 2007. Meteor: an automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 228–231, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Liu and Gildea2005] Ding Liu and Daniel Gildea. 2005. Syntactic features for evaluation of machine translation. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 25–32.

[Lo et al.2012] Chi-kiu Lo, Anand Karthik Tumuluru, and Dekai Wu. 2012. Fully automatic semantic mt evaluation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 243–252, Montréal, Canada, June. Association for Computational Linguistics.

[Mehay and Brew2007] Dennis Mehay and Chris Brew. 2007. BLEUTRE: Flattening Syntactic Dependencies for MT Evaluation. In *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*.

[Och2003] F.J. Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 160–167. Association for Computational Linguistics.

[Owczarzak et al.2007] Karolina Owczarzak, Josef van Genabith, and Andy Way. 2007. Labelled dependencies in machine translation evaluation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 104–111, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Papineni et al.2002] K. Papineni, S. Roukos, T. Ward, and W.J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

[Porter2001] Martin F Porter. 2001. Snowball: A language for stemming algorithms.

[Shannon2001] Claude Elwood Shannon. 2001. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55.

[Snover et al.2006] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

[Zhu et al.2010] Junguo Zhu, Muyun Yang, Bo Wang, Sheng Li, and Tiejun Zhao. 2010. All in strings: a powerful string-based automatic mt evaluation metric with multiple granularities. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 1533–1540, Stroudsburg, PA, USA. Association for Computational Linguistics.